COMPARING PROBABILITY FORECASTERS: BASIC BINARY CONCEPTS AND MULTIVARIATE EXTENSIONS

by

Morris H. DeGroot

and

Stephen E. Fienberg

Department of Statistics

Carnegie-Mellon University

September, 1983

Technical Report No. 306

This research was supported in part by the National Science Foundation under grant SES-8207295 and by the Office of Naval Research under contract N00014-80-C-0637.

1. INTRODUCTION

In the applied forecasting literature much attention has been lavished on questions about the evaluation of probability forecasts, and the subjectivist view of probability has been invoked to aggregate probability forecasts over a diverse set of events or statements, (e.g. see Fischhoff and McGregor, 1982). One criterion invoked in such evaluations is that of calibration: a set of statements or events is considered and we ask if x percent of those assigned probability x of being correct prove to be correct, for each value of x. From this perspective, weather forecasters generally have been found to perform well, (Murphy and Winkler, 1974, 1977). What is especially helpful in the evaluation of such probability forecasters is that they make forecasts about a long sequence of events (e.g. rain on a given day), and thus it makes sense to think about probability functions associated with the forecasts. In this paper we focus on a criterion for comparing forecasters, refinement, which goes beyond that of calibration, (see the related discussion in Winkler, 1982).

The formal setting we consider is the same as that presented in DeGroot and Fienberg (1982, 1983). Consider two forecasters who at the beginning of each period n in a sequential process (n = 1,2,...) must independently specify their subjective probabilities that a particular event A_n will occur during the period. Assume that each forecaster in specifying the probability of A_n is aware of the values of various variables that are potentially relevant to the occurrence of A_n, including which of the previous events A_1, A_2, ..., A_{n-1} have actually occurred. We wish to compare the two forecasters on the basis of their subjective probabilities of the events A_1, A_2, ..., A_n and the subsequent observation of exactly which of those events occurred, for large values of n.

It is possible to think of our forecasters as economists who at the beginning of each period must specify their probabilities that the value of a particular commodity will rise during the period, or as medical diagnosticians who specify their probabilities that a patient has a particular disorder (e.g. see Habbema, Hilden, and Bjerregaard, 1978, and Hilden, Habbema and Bjerregaard, 1978a, 1978b). As in DeGroot and Fienberg (1982, 1983), however, we present the basic results of the first four sections of this paper in the context of two weather forecasters. Day after day, they must specify their subjective probabilities, x and y, that there will be at least a certain amount of rain at some given location during a specified time period in the day, and we refer to the occurrence of this well-specified event as "rain." We make the assumption that x and y are restricted to a given finite set of values $0 = x_0 < x_1 < \ldots < x_k = 1$, and we let X denote the set $\{x_0, x_1, x_2, \ldots, x_k\}$.

In Section 2, we present the basic notation and formal definitions of the concepts of *calibration* and *refinement*. Then we go on to describe the relationship between these concepts and the classical concept of *sufficiency*, and we summarize various results on when one well-calibrated forecaster is at least as refined as another given in DeGroot and Fienberg (1982, 1983) and DeGroot and Eriksson (1983). In Section 3, we present two concepts introduced by Vardeman and Meedan (1983), *semi-calibration* and *rain (or dry)-domination*, which impose different restrictions on the probability distributions of the forecasters than do calibration and sufficiency or refinement, and we discuss how these different concepts are related.

In Section 4, we introduce the notion of strictly proper scoring rules and describe their use in comparing forecasters. We give a proof of a basic partitioning result for strictly proper scoring rules, described earlier in DeGroot and Fienberg (1983). We also give a result on the relationship between Schur-convex measures of the quality of forecasters and strictly proper scoring rules presented in DeGroot and Eriksson (1983).

In Section 5, we turn to the multivariate or vector-probability forecasting situation in which the events of interest have three or more possible outcomes. For example, in the weather context the event may be tomorrow's maximum temperature and the possible outcomes may be grouped into 5°C intervals. The forecasters are required to announce their subjective probabilities associated with each of the $s \geq 3$ possible outcomes.

2. CALIBRATION AND REFINEMENT

We begin by considering two forecasters, A and B, and we define the joint probability function, $g(x,y,\theta)$, for the random variables associated with the ith event E_i in a sequence where x and y are forecaster A's and B's subjective probabilities of rain, respectively, and $\theta = 1$ if the outcome is rain or $\theta = 0$ otherwise. The overall performance of forecaster A can be characterized by two functions defined on X:

(i) the probability function

$$\nu_A(x) = \sum_{y \in X} \sum_{\theta} g(x,y,\theta)$$

which gives the probabilities or relative frequencies with which forecaster A makes each possible prediction x.

(ii) the conditional probability function

$$\rho_A(x) = \sum_{y \in X} g(x,y,1)/\nu_A(x)$$

which give the conditional probability or relative frequency of rain given forecaster A's specific prediction x.

The functions $\nu_B(x)$ and $\rho_B(x)$ for forecaster B can be defined similarly. Finally the long run frequency of rain is

$$\mu = \sum_{x \in X} \sum_{y \in X} g(x,y,1).$$

A forecaster is said to be *well-calibrated* if $\rho(x) = x$ for all $x \in X$ such that $\nu(x) \neq 0$. Thus a forecaster is well-calibrated if the forecaster's predictions can be accepted at face value, i.e. given that the forecaster predicts x, the conditional probability of rain is x. DeGroot and Eriksson (1983) give three different interpretations for the quantities $\nu(x)$ and $\rho(x)$, based on (i) limiting or theoretical values for an infinite sequence of days, (ii) the actual relative frequencies for a finite sequence of n days, and (iii) subjective probabilities of an observer who is comparing the forecasters. We proceed using interpretation (iii), although we sometimes use language that is associated with (i).

For various reasons, being well-calibrated is usually regarded as a desirable characteristic of a forecaster. For example, Pratt (1962) and Dawid (1982) show that a probability forecaster who is coherent in the sense of de Finetti (1937) must be calibrated almost surely. However, as Murphy and Winkler (1977), Dawid (1982), and DeGroot and Fienberg (1983) note, even if a forecaster is well-calibrated his predictions are not necessarily accurate in all respects nor are they necessarily of much use to anyone. For example, the forecaster whose prediction on each day is μ is well-calibrated but clearly useless as a forecaster once we know μ.

Consider two well-calibrated forecasters A and B whose predictions are characterized by their probability functions ν_A and ν_B. In DeGroot and Fienberg (1982, 1983), we introduce a concept of refinement which induces a partial ordering on the class of all well-calibrated forecasters. This concept is defined as follows:

A *stochastic transformation* $h(y|x)$ is a non-negative function defined on $X \times X$ such that

$$\sum_{y \in X} h(y|x) = 1 \text{ for every } x \in X. \tag{2.1}$$

Forecaster A is said to be *at least as refined* as forecaster B if there exists a stochastic transformation $h(y|x)$ such that

$$\sum_{x \in X} h(y|x)\nu_A(x) = \nu_B(y) \qquad \text{for } y \in X. \tag{2.2}$$

$$\sum_{x \in X} h(y|x)x\nu_A(x) = y\nu_B(y) \qquad \text{for } y \in X. \tag{2.3}$$

Following DeGroot and Eriksson (1983) we denote this relationship by the symbols $A \succsim B$. The stochastic transformation here plays the role of an auxiliary randomization which we could use to generate predictions with distribution $\nu_B(y)$ from A's predictions, as in (2.2). Equation (2.3) is required in the definition to ensure that the predictions generated by this process are well-calibrated.

If $A \succsim B$ and the probability functions ν_A and ν_B are not identically equal, then A is said to be *more refined* than B. We denote this relationship by the symbols $A \succ B$.

The relationship $A \succsim B$ is both reflexive and transitive, and induces a partial ordering (but not a total ordering) among well-calibrated forecasters. In these terms, the forecaster who makes the same prediction μ each day is *least-refined* in the sense that any other well-calibrated forecaster is at least as refined as he is. It is possible that μ is not one of the allowable predictions $x_0, x_1, \ldots, x_k$. In that case, there is a value of i (i = 0,1,...,k-1) such that $x_i < \mu < x_{i+1}$, and DeGroot and Fienberg (1982) show that a forecaster who uses only the values x_i and x_{i+1} as his predictions in such a way that he is well calibrated will now be least-refined. At the other extreme, the forecaster whose prediction each day is either $x = 0$ or $x = 1$ and who is always correct is *most-refined* in the sense that he is more refined than any other well-calibrated forecaster.

For any well-calibrated forecaster it must be true that

$$\sum_{x \epsilon X} x\nu(x) = \mu \ . \tag{2.4}$$

Since every distribution with mean μ *can be thought of* as the $\nu(x)$ of some well-calibrated forecaster, the comparison of well-calibrated forecasters using the relationship $\succsim$ is equivalent to the problem of the comparison of all distributions on X with a given mean μ.

The left-hand side of expression (2.3) resembles the form of a conditional expectation. This observation leads to the following result:

Theorem 1 (DeGroot and Eriksson, 1983): The relationship $A \succsim B$ is satisfied if and only if there exist discrete random variables X and Y such that the marginal probability distribution of X is ν_A, the marginal probability distribution of Y is ν_B, and $E(X \mid Y) = Y$.

We temporarily step back and consider the comparison of two arbitrary forecasters A and B, who are not necessarily well calibrated. For any given forecaster, let $f(x \mid \theta)$ denote the conditional probability function of the forecaster's predictions given θ. Thus, $f(x \mid 1)$ can be regarded as the frequency function of the forecaster's predictions on days when rain actually occurs, and $f(x \mid 0)$ as the frequency function on days when rain does not occur. It follows that for $x \epsilon X$,

$$f(x\,|\,1) = \rho(x)\nu(x)/\mu \ , \tag{2.5}$$

$$f(x\,|\,0) = \{1 - \rho(x)\}\nu(x)/(1 - \mu) \ . \tag{2.6}$$

Thus we can also use the probability function $f(x\,|\,\theta)$ for $\theta = 1$ and $\theta = 0$ to characterize a forecaster's predictions. Let forecasters A and B *have conditional probability functions* $f_A(x\,|\,\theta)$ and $f_B(y\,|\,\theta)$. Following Blackwell's (1951, 1952) work on the comparison of experiments we say that A is *sufficient* for B if there exists a stochastic transformation $h(y\,|\,x)$ such that

$$\sum_{x\epsilon X} h(y\,|\,x)\ f_A(x\,|\,\theta) = f_B(y\,|\,\theta) \qquad \text{for } y\,\epsilon\, X \text{ and } \theta = 0,1 \ . \tag{2.7}$$

Using the relationships (2.5) and (2.6) we can now prove

Theorem 2 (DeGroot and Fienberg, 1982): Consider two forecasters A and B whose predictions are characterized by $\nu_A(x)$, $\rho_A(x)$, $\nu_B(x)$, and $\rho_B(x)$. Then A is sufficient for B if and only if there exists a stochastic transformation h such that

$$\sum_{x\epsilon X} h(y\,|\,x)\nu_A(x) = \nu_B(y) \qquad \text{for } y\,\epsilon\, X. \tag{2.8}$$

and

$$\sum_{x\epsilon X} h(y\,|\,x)\rho_A(x)\nu_A(x) = \rho_B(y)\nu_B(y) \qquad \text{for } y\,\epsilon\, X. \tag{2.9}$$

Now if we again restrict attention to well calibrated forecasters we get the following corollary:

Corollary 1: Consider two well-calibrated forecasters A and B. Then $A \succsim B$ if and only if A is sufficient for B.

Thus we can use results from the comparison of experiments to compare two well-calibrated probability forecasters A and B. For any well-calibrated forecaster, let F denote the distribution function corresponding to the probability function ν, i.e.

$$F(t) = \sum_{\{x:x\,\epsilon\, X, x \le t\}} \nu(x) \qquad \text{for } 0 \le t \le 1 \ . \tag{2.10}$$

Theorem 3 (DeGroot and Eriksson, 1983): The relationship $A \succsim B$ is satisfied if and only if

$$\int_0^s F_A(t)dt \ge \int_0^s F_B(t)dt \qquad \text{for all } 0 \le s \le 1 \ . \tag{2.11}$$

The relationship (2.11) is one of several equivalent definitions of *second-degree stochastic* *dominance* (e.g. see Fishburn and Vickson, 1978), and we see that it is equivalent to the relationship A $\succsim$ B. This leads to yet another equivalence:

Theorem 4 (Hardy, Littlewood, and Polya, 1919, 1934): The relationship A $\succsim$ B is satisfied if and only if, for every continuous, convex function c defined on the unit interval,

$$\Sigma_{x \in X} c(x) \nu_A(x) \geq \Sigma_{x \in X} c(x) \nu_B(x) \; . \tag{2.12}$$

Finally, there is a simplification of the second-order stochastic dominance relationship that can be expressed only in terms of the probability functions ν_A and ν_B.

Theorem 5 (DeGroot and Fienberg, 1982): The relationship A $\succsim$ B is satisfied if and only if

$$\Sigma_{i=0}^{j-1} (x_j - x_i)\{\nu_A(x_i) - \nu_B(x_i)\} \geq 0 \qquad \text{for } j = 1, \ldots, k-1 \; . \tag{2.13}$$

If at least one inequality in (2.13) is strict, then A $\succ$ B.

By a direct application of this theorem, we can gain insight into the refinement relationship through the following Corollary:

Corollary 2: Suppose A and B are well calibrated forecasters such that

$$\nu_B(x_i) \geq \nu_A(x_i) \qquad i = 1, \ldots, k-1 \tag{2.14}$$

and $\nu_B(x_0) = \nu_B(x_k) = 0$. Then A $\succsim$ B. If one of the inequalities in (2.14) is strict, then A $\succsim$ B.

Thus we see that, in a rough sense, A is more refined than B if A spreads his probabilities into the extremes (in this case to 0 to 1) more than B.

As we have seen from (3.5), in order to determine whether A is sufficient for B, we need only know the marginal probability functions $f_A(x|\theta)$ and $f_B(x|\theta)$ of A and B separately. However, in order to determine whether forecaster A is sufficient for both himself and forecaster B together, we must know the joint probability function $\nu(x,y) = g(x,y,1) + g(x,y,0)$ of the predictions x and y of A and B, as well as the conditional probability of rain given the two predictions x and y:

$$\rho(x,y) = g(x,y,1) \Big/ \nu(x,y) \ . \tag{2.15}$$

Then we can use the following result:

Theorem 6 (DeGroot and Fienberg, 1983): Forecaster A is sufficient for the pair of forecasters (A,B) if and only if

$$\rho(x,y) = \rho_A(x) \qquad \text{for } x \in X \text{ and } y \in X \tag{2.16}$$

If follows that, if A sufficient for (A,B), then A is sufficient for B. The converse, however, is not necessarily true.

Suppose now that neither A nor B is sufficient for the other. It becomes natural to ask if we can do better than A or B by using only their predictions. One way to try to do this is to choose A's prediction with probability α and B's with probability $1-\alpha$. This results in a "new forecaster" whom we label as M(α). If both A and B are well-calibrated then it is straightforward to show that M(α) is also well-calibrated. Furthermore, it follows directly from Theorem 3 that M(α) $\succeq$ A if and only if B $\succeq$ A, and that M(α) $\succeq$ B if and only if A $\succeq$ B. Thus, randomly mixing two well-calibrated forecasters does not allow us to improve upon either of them in the refinement sense.

The second way to use the predictions of two forecasters is to average them (or more generally take linear combinations). Unfortunately, if A and B are well-calibrated it does not necessarily follow that the average of A's and B's predictions will be well-calibrated. This is most easily seen if we average the predictions of the most-refined well-calibrated forecaster, who always correctly forecasts 0 or 1, and the predictions of the least-refined well-calibrated forecaster, who always forecasts μ. Thus if we wish to improve upon A and B by averaging their predictions, there is the potential loss of calibration. Moreover, even if the result of averaging A's and B's forecasts is well-calibrated, it remains unclear whether the distribution of the average can be more spread than A's and B's distributions.

3. RESTRICTED COMPARISONS

As in many other statistical decision problems, there are essentially two different types of error that the forecaster might make. He might predict a high probability of rain on a day when it does not rain or he might predict a low probability of rain on a day when it does rain. Some forecasters may control one type of error better than the other. We may thus wish to consider how to compare two forecasters separately for days on which it rains and for days on which it does not (i.e. dry days). This leads to two notions of dominance, introduced by Vardeman and Meeden (1983), which are both forms of *first-degree stochastic dominance* (again see Fishburn and Vickson, 1978).

We say that forecaster A *rain dominates* forecaster B if

$$\Sigma^j_{i=0} f_A(x_i|1) \le \Sigma^j_{i=0} f_B(x_i|1) \qquad \text{for } j = 0,1,2,\ldots, k. \qquad (3.1)$$

dry dominates forecaster B if

$$\Sigma^j_{i=0} f_A(x_i|0) \ge \Sigma^j_{i=0} f_B(x_i|0) \qquad \text{for } j = 0,1,2,\ldots, k, \qquad (3.2)$$

and *dominates* forecaster B if both (3.1) and (3.2) hold. Condition (3.1) says that on rainy days, A's predictions have a distribution which is stochastically larger than the distribution of B's predictions, and Condition (3.2) says that for dry days A's predictions are stochastically smaller than B's predictions.

The relationships of rain domination, dry domination, and domination, like that of refinement (or sufficiency), each induce their own partial ordering amongst forecasters. They also provide alternative ways of demonstrating refinement, as the following result shows:

> *Theorem 7* (Vardeman and Meeden, 1983): For two well-calibrated forecasters A and B, if either (i) A rain dominates B or (ii) A dry dominates B, then $A \succsim B$.

The relationship of domination is quite stringent and thus it does not need to be combined with a condition as strong as calibration to imply sufficiency. Following Vardeman and

Meeden, we say that a forecaster is *semi-calibrated* if $\rho(x)$ is nondecreasing in x for those values of x with $\nu(x) > 0$.

Theorem 8 (Vardeman and Meeden, 1983): If forecasters A and B are both semi-calibrated and A dominates B, then A is sufficient for B.

Vardeman and Meeden (1983) go on to use the concepts of domination and semi-calibration, along with calibration and refinement, to make comparisons between Bayesian forecasters who use stationary n-step Markov chain representations for the sequence of outcomes θ.

4. STRICTLY PROPER SCORING RULES

It has often been suggested in the statistical literature that a forecas[illegible] predictions over a sequence of days can be evaluated by the use of a *scoring rule* whi[illegible]ssigns a numerical value, or score, each day based on the forecaster's prediction x and the [illegible]ation of whether or not rain occurred, i.e. the observation of θ. One property of the use of such rules, when the forecaster attempts to maximize the expectation of this score, is that if the forecaster's predictions are not restricted to be probabilities, then there is a known transform of the values of x to values which are probabilities (Lindley, 1982). For the class of proper scoring rules described below, the values of x must themselves be probabilities. There is little reason, however, to believe that a forecaster will want to maximize his expected overall score (e.g. see the discussion on this point in DeGroot and Fienberg, 1983, and in Stael von Holstein, 1970).

Our interest in scoring rules in the context of comparing forecasters is somewhat different. Since we know that the relationship $A \succeq B$ induces only a partial ordering on the class of well-calibrated forecasters, we wish to assign a measure of quality $m(\nu)$ to the probability function ν of every well-calibrated forecaster in order to obtain a total ordering of this class. The values $m(\nu)$ should be assigned in such a way that the better the forecaster, the higher his measure of quality will be. It is natural to interpret this requirement to mean that if $A \succeq B$, then $m(\nu_A) \geq m(\nu_B)$, with strict inequality unless the probability functions ν_A and ν_B are identical. Indeed, we need not restrict attention to only well-calibrated forecasters. It is convenient to arrive at the measure $m(\cdot)$ for the comparison of well-calibrated forecasters from the more general approach of the use of scoring rules which are applicable to all forecasters.

We begin by considering an arbitrary scoring rule. Suppose that the forecaster's prediction is x and rain occurs. Then the forecaster receives a score $g_1(x)$. If rain does not occur he receives a score $g_2(x)$. Since we assume that the forecaster desires to maximize his score, it is reasonable to assume that $g_1(x)$ is an increasing function of x and that $g_2(x)$ is a decreasing function of x. If the forecaster's actual subjective probability of rain on a particular day is p and he makes the prediction x, then his expected score is

$$pg_1(x) + (1-p)g_2(x) \; . \qquad (4.1)$$

A *proper scoring rule* is one for which expression (4.1) is maximized when $x = p$. A *strictly proper scoring rule* is one for which $x = p$ is the *only* value of x that maximizes expression (4.1). An interesting discussion of these rules, with historical references, is given by Staël von Holstein (1970, Sec. 3.2).

Examples of strictly proper scoring rules include the quadratic rule (Brier, 1950; de Finetti, 1962, 1965) with $g_1(x) = -(x-1)^2$ and $g_2(x) = -x^2$, and the logarithmic rule (Good, 1952) with $g_1(x) = \log x$ and $g_2(x) = \log (1-x)$. Both of these examples have the symmetry property, $g_1(x) = g_2(1-x)$, but this is not a requirement of strictly proper rules. An example of an *improper* scoring rule is the exponential with $g_1(x) = e^{x/2}$ and $g_2(x) = e^{-x/2}$. Here the values that maximize the expected score are $x = \log\{(1-p)/p\}$, i.e. the log-odds.

If a proper scoring rule is used for all of a forecaster's predictions then we get an overall score S for the forecaster. Among all days on which the forecaster's prediction is x, the score will be $g_1(x)$ with relative frequency $\rho(x)$ and $g_2(x)$ with relative frequency $1-\rho(x)$. Since the relative frequency of the prediction x for the forecaster is $\nu(x)$, we have that

$$S = \sum_{x \in X} \nu(x) \; \{\rho(x)g_1(x) + [1-\rho(x)] \; g_2(x)\} \; . \qquad (4.2)$$

We now come to the major result of this Section, which shows that this overall score (4.2) can be partitioned into a component which is a measure of calibration and a component which measures sufficiency or refinement.

Theorem 9 (DeGroot and Fienberg, 1983): If a forecaster's predictions are characterized by the functions $\nu(x)$ and $\rho(x)$, and if a proper scoring rule is specified by the functions $g_1(x)$ and $g_2(x)$, then the forecaster's overall score S can be expressed in the form $S = S_1 + S_2$, where

$$S_1 = \sum_{x \in X} \nu(x) \, [\rho(x) \, [g_1(x) - g_1\{\rho(x)\}] + \{1-\rho(x)\} \, [g_2(x) - g_2\{\rho(x)\}]], \tag{4.3}$$

$$S_2 = \sum_{x \in X} \nu(x)\varphi\{\rho(x)\}, \tag{4.4}$$

and

$$\varphi(t) = tg_1(t) + (1-t)g_2(t) \qquad \text{for } 0 \le t \le 1 . \tag{4.5}$$

If the scoring rule is strictly proper then $\varphi(t)$ is strictly convex and S_1 attains its maximum value only when $\rho(x) = x$ for every value of x such that $\nu(x) > 0$.

Proof: It can be verified directly that if S_1 and S_2 are given by (4.3) to (4.5) then $S_1 + S_2$ is given by expression (4.2) for S. Thus the first part of the theorem is established. Now suppose that the scoring rule is strictly proper. To see that $\varphi(t)$ is convex, note that

$$\varphi(t) = \max_{0 \le x \le 1} \, [tg_1(x) + (1-t)g_2(x)]. \tag{4.6}$$

In (4.6), $\varphi(t)$ is represented as the maximum of a family of linear functions of t. Hence, $\varphi(t)$ is convex.

If $\varphi(t)$ is not strictly convex, then it contains at least one linear segment. This means that there must be one particular linear function that yields the maximum value in (4.6) for all the values of t in some interval. But that is impossible, because we know that at any particular value $t = t_0 (0 \le t_0 \le 1)$ the maximum in (4.6) is attained uniquely by the linear function $tg_1(t_0) + (1-t)g_2(t_0)$. Hence, $\varphi(t)$ must be strictly convex.

Finally, since the scoring rule is strictly proper we know that

$$\rho(x)g_1(x) + [1-\rho(x)]g_2(x) \le$$

$$\rho(x)g_1[\rho(x)] + [1-\rho(x)]g_2[\rho(x)], \tag{4.7}$$

with strict inequality unless $x = \rho(x)$. Hence, it can be seen from (4.3) that S_1 will be negative unless $\rho(x) = x$ for every value of x such that $\nu(x) > 0$, in which case S_1 will be equal to zero. ■

As we noted above, S_1 is a measure of the forecaster's calibration, which is zero only for a well-calibrated forecaster and negative otherwise. The component S_2 provides us with our sought-after measure to give a total ordering for well-calibrated forecasters:

$$m(\nu) = \Sigma_{x \epsilon X}\ \phi(x)\nu(x), \tag{4.8}$$

which is S_2 with $\rho(x) = x$. Since $\phi(x)$ is strictly convex, Theorem 4 implies that if $A \succsim B$, then $m(\nu_A) \geq m(\nu_B)$, with strict inequality unless ν_A and ν_B are identical.

DeGroot and Eriksson (1983) note that there is a direct relationship between the total ordering provided by the measure m and the concept of Schur-convexity which plays an important role elsewhere in statistics (Marshall and Olkin, 1979). Consider a function m defined on the class of all probability distributions ν over X that have a given mean μ. Then m is said to be strictly *Schur-convex* if $m(\nu_A) \geq m(\nu_B)$ whenever the relation (2.11) is satisfied with strict inequality unless $\nu_A = \nu_B$. The following result now follows directly from Theorem 4.

> *Theorem 10* (DeGroot and Eriksson, 1983): Consider a strictly proper scoring rule based on the functions g_1 and g_2, and suppose that a measure of quality m is defined by (4.8) and (4.5). Then m is strictly Schur-convex.

Suppose now that we know the functions $\nu(x)$ and $\rho(x)$ that characterize a particular forecaster's predictions. Is it possible for us to use his predictions, and no other relevant meteorological information, to make our own predictions and to attain a larger value of the score S than the forecaster himself? The following argument generalizes the one given in DeGroot and Fienberg (1983) for the quadratic scoring rule (see also Schervish, 1983).

The forecaster's score is given by expression (4.2). In order for us to make our predictions, we must choose a stochastic transformation $h(x|y)$ as follows. If the forecaster's prediction on a given day is y, then we choose our prediction at random from X in accordance with the conditional distribution $h(x|y)$. With this procedure, our predictions are characterized by the functions

$$\nu_0(x) = \sum_{y\in X} h(x|y)\nu(y) \ , \tag{4.9}$$

$$\rho_0(x) = \sum_{y\in X} h(x|y)\rho(y)\nu(y)/\nu_0(x) \ . \tag{4.10}$$

It follows from expressions (4.2), (4.9), and (4.10) after some algebra, that our score is

$$S_0 = \sum_{y\in X} \nu(y) \sum_{x\in X} [\rho(y)g_1(x) + (1-\rho(y))g_2(x)] \ h(x|y) \ . \tag{4.11}$$

For each fixed value of y, the summation over x in expression (4.11) yields a weighted average of the quantities

$$\rho(y)g_1(x) + \{1-\rho(y)\} \ g_2(x) \ , \tag{4.12}$$

with weights given by the conditional probabilities $h(x|y)$. Thus to maximize the weighted average, we choose the conditional distribution $h(x|y)$ to put all the probability on the value of x that maximize expression (4.12). If $\rho(y)\in X$ the maximizing value is $x = \rho(y)$, and we make the forecaster well-calibrated. With this choice, our value of S_2 remains the same as that of the original forecaster, but our value of S_1 is now increased to 0. If $\rho(x)\notin X$, then we come as close to the maximum of expression (4.12) as possible, by setting x equal to the permissible value close to $\rho(y)$ that maximizes (4.12), i.e. we make the forecaster *almost* well-calibrated. Formally, a forecaster is said to be *almost well-calibrated* (relative to the strictly proper scoring rule defined by g_1 and g_2) if for each point $y\in X$ such that $\nu(y) > 0$, the expression (4.12) is maximized over the points $x\in X$ when $x = y$. Following Schervish (1983), if we take an arbitrary forecaster B we refer to a second forecaster A who uses this concentrated function $h(x|y)$ to transform the predictions of B into his own as "the almost calibrated version of B." Then our result is:

Theorem 11. Consider a strictly proper scoring rule. Let B be any forecaster and let A be the almost calibrated version of B. If B is not almost well calibrated, then

A has a strictly larger score than B.

This theorem can be viewed as providing motivation for the idea of recalibrating forecasters suggested by Lindley, Tversky, and Brown (1979).

5. COMPARING MULTIVARIATE FORECASTERS

We now turn to a consideration of forecasting events with s > 2 outcomes (e.g. a set of temperature ranges). In such settings the probability forecaster specifies a vector of probabilities $\mathbf{x} = (x_1, x_2, \ldots, x_s)$, restricted to a finite set $X^{(s)}$ of values lying in the (s-1)-dimensional simplex, i.e. $x_i \geq 0$ and $\Sigma_{i=1}^{s} x_i = 1$. If the conditional probabilities of the s outcomes given the prediction x are represented in vector form by $\rho(\mathbf{x}) = [\rho_1(\mathbf{x}), \rho_2(\mathbf{x}), \ldots, \rho_s(\mathbf{x})]$, then the multivariate forecaster is well-calibrated if $\rho(\mathbf{x}) = \mathbf{x}$ for all $\mathbf{x} \in X^{(s)}$. Note that this well-calibrated multivariate forecaster is also well-calibrated, in the sense of Section 2, for each binary problem formed by combining the s outcomes into two groups; however, a forecaster who is "marginally" well-calibrated for predicting "rain" or "no rain" may no longer be well-calibrated when "rain" is divided into two or more possible outcomes.

More formally, let $\mathbf{x} = (x_1, \ldots, x_s)$ and $\rho(\mathbf{x}) = [\rho_1(\mathbf{x}), \ldots, \rho_s(\mathbf{x})]$. Furthermore, let $\mathcal{I} = \{I_1, \ldots, I_k\}$ represent a partition of the set {1,...,s} into k nonempty, mutually exclusive, and exhaustive sets $I_1, \ldots, I_k$. Then a forecaster is said to be *marginally well-calibrated with respect to the partition* $\mathcal{I}$ if

$$\Sigma_{i \in I_j} \rho_i(\mathbf{x}) = \Sigma_{i \in I_j} x_i \quad \text{for } j = 1, \ldots, k \text{ and } \mathbf{x} \in X^{(s)}. \tag{5.1}$$

We can also focus on a particular set of the partition $\mathcal{I}$, say I_n, and define

$$\rho_i(\mathbf{x}, I_n) = P(\theta = i \mid \theta \in I_n, \text{forecast } \mathbf{x}). \tag{5.2}$$

Then we can say that a forecaster is *conditionally well-calibrated given the set* I_n if

$$\rho_i(\mathbf{x}, I_n) = \frac{x_i}{\Sigma_{j \in I_n} x_j}, \quad i \in I_n. \tag{5.3}$$

Moreover, because being well calibrated in the multivariate sense is a demanding requirement, we might also want to know if a forecaster is well-calibrated for some but not necessarily all values of x. Let $\mathcal{X}_0^{(s)}$ denote a proper subset of $\mathcal{X}^{(s)}$. Then we say that a forecaster is *partially well-calibrated on the subset* $\mathcal{X}_0^{(s)}$ if $\rho(\mathbf{x}) = \mathbf{x}$ for $\mathbf{x} \in \mathcal{X}_0^{(s)} \subset \mathcal{X}^{(s)}$. We can now combine these notions of partial, conditional, and marginal calibration in various ways. (In particular, we note that the concept of conditional calibration suggested in DeGroot and Fienberg (1982) is in fact a combination of conditional and partial calibration as defined here.) We also consider an extension of semi-calibration, introduced in Section 3, to the multivariate setting in a special case at the end of this section.

For well-calibrated multivariate forecasters, we can define the concept of refinement by means of a multivariate stochastic transformation $h(\mathbf{x}\,|\,\mathbf{y})$. Consider two well-calibrated forecasters characterized by their probability functions ν_A and ν_B. Then we say that A is at least as refined as B if there exists a stochastic transformation h such that:

$$\sum_{\mathbf{x} \in \mathcal{X}^{(s)}} h(\mathbf{y}\,|\,\mathbf{x})\mathbf{x}^T \nu_A(\mathbf{x}) = \mathbf{y}^T \nu_B(\mathbf{y}) \text{ for } \mathbf{y} \in \mathcal{X}^{(s)} . \tag{5.4}$$

Note that the analogue of equation (2.2), i.e.

$$\sum_{\mathbf{x} \in \mathcal{X}^{(s)}} h(\mathbf{y}\,|\,\mathbf{x}) \nu_A(\mathbf{x}) = \nu_B(\mathbf{y}) \text{ for } \mathbf{y} \in \mathcal{X}^{(s)} . \tag{5.5}$$

is automatically satisfied by summing the s equations in expression (5.4). Furthermore, we can immediately define concepts of marginal refinement with respect to a partition $\mathcal{I}$. The concept of conditional refinement given the set I_n which also appears to be immediate (in a definitional sense) is, however, problematic as it involves conditioning of the vector x on $\theta \in I_n$. These conditional predictions have no operational meaning, because we cannot define them only in terms of the probability distribution ν. Similarly, the concept of partial refinement on the subset $\mathcal{X}_0^{(s)} \subset \mathcal{X}^{(s)}$ also is problematic since two different forecasters typically place different amounts of subjective probability on the set $\mathcal{X}_0^{(s)}$.

At any rate, Theorem 2 and Corollary 1 from Section 2 carry over directly from the binary case, i.e., forecaster A is sufficient for forecaster B if and only if there exists an appropriate stochastic transformation, $h(\mathbf{x}\,|\,\mathbf{y})$. Moreover, suppose we define a multivariate scoring rule, $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), \ldots, g_s(\mathbf{x})]$. If the forecaster's actual subjective probability is p and he makes the prediction x, his expected score is

$$\Sigma p_i g_i(x). \qquad (5.6)$$

The scoring rule is strictly proper if expression (5.6) is maximized if and only if x = p. Then there is a direct multivariate analogue to Theorem 9 of Section 3, i.e. every strictly proper scoring rule can be partioned into two components, one of which is zero if the forecaster is well-calibrated and the other of which is a measure of refinement giving a total ordering for well-calibrated forecasters.

The following results are also as expected:

Theorem 12. If a multivariate forecaster A is well-calibrated, (i) A is also marginally well-calibrated with respect to all possible proper partitions $\mathcal{I}$ of $\{1,2,\ldots,s\}$, (ii) A is conditionally well-calibrated given the set $I_n \subset \{1,2,\ldots,s\}$, and (iii) A is partially well-calibrated on all proper subsets $X_0^{(s)} \subset X^{(s)}$.

Theorem 13. If A and B are well-calibrated multivariate forecasters, and A is at least as refined as B, then A is also marginally at least as refined as B with respect to all possible proper partitions $\mathcal{I}$.

We have, as yet, been unable to provide a collection of refinement conditions for dichotomies which imply multivariate refinement. Nor have we been able to prove a directly verifiable set of conditions analogous to Theorem 5 of Section 3. We can, however, give multivariate versions of Theorems 1 and 4 by reformulating results of Blackwell (1951, 1953), Sherman (1951), Stein (in unpublished lecture notes), and Strassen (1965).

Theorem 14. Consider two well-calibrated forecasters A and B. Then A is at least as refined as B if and only if there exist discrete random variables x and y, defined on the (s-1)-dimensional simplex, such that the marginal probability distribution of X is ν_A, the marginal probability distribution of Y is ν_B, and $E(X \mid Y) = Y$.

Theorem 15. Consider two well-calibrated forecasters A and B. Then A is at least as refined as B if and only if, for every continuous convex function c(x) defined on the (s-1)-dimensional simplex

$$\sum_{x \in X^{(s)}} c(x)\nu_{A}(x) \geq \sum_{x \in X^{(s)}} c(x)\nu_{B}(x) . \qquad (5.7)$$

We note that expression (5.7) in Theorem 15 is, in our problem, the same as the condition used by Fishburn and Vickson (1978) for their definition of multivariate second-degree stochastic dominance. They also suggest the application of standard feasibility tests of linear programming to determine the existence of the stochastic transformation which we use to define refinement.

Furthermore, we have the following direct multivariate extension of a result presented in DeGroot and Eriksson (1983).

Theorem 16. Consider two well-calibrated forecasters A and B. Then A is at least as refined as B if and only if there exists a stochastic transformation η such that

$$\sum_{x \in X^{(s)}} x^{T} \, \eta(x \mid y) = y^{T} \text{ for } y \in X^{(s)} . \qquad (5.8)$$

Proof: Suppose that A is at least as refined as B, and let h be a stochastic transformation satisfying (5.4). If we define

$$\eta(x \mid y) = \frac{h(y \mid x)\nu_{A}(x)}{\nu_{B}(y)} \qquad (5.9)$$

whenever $\nu_{B}(y) > 0$, and define $\eta(x \mid y)$ arbitrarily of $\nu_{B}(y) = 0$, then (5.8) follows directly from (5.4). Conversely, suppose that (5.8) is satisfied for some η and define the stochastic transformation h by (5.9). [Note that $h(y \mid x)$ may be defined arbitrarily if $\nu_{A}(x) = 0$.] Then (5.4) follows directly from (5.8). ■

A stochastic transformation η satisfying expression (5.8) is known in the economics literature as a *mean-preserving spread* (see, e.g., Rothschild and Stiglitz, 1970, 1973).

An interesting version of the multivariate setting results when the probability of outcome i given that the forecaster predicts x depends only on the forecaster's subjective probability x_i

for outcome i. (This is clearly true if the forecaster is well-calibrated). We say that the forecaster is *local* if

$$\rho_i(x) = \rho_i(x_i) \qquad i = 1,2,\dots,s. \tag{5.10}$$

If the functions $\rho_i(\cdot)$ are monotonically increasing, then the forecaster is marginally semi-calibrated, in the sense of Section 3. A special case of locality is linearity, and an interesting question arises: Under what conditions on $\mathcal{X}^{(s)}$ and the ρ_i's is a multivariate forecaster being local equivalent to

$$\rho_i(x) = b_0 x_i + b_i \qquad i = 1,2,\dots,s. \tag{5.11}$$

where $b_0, b_1, \dots, b_s \geq 0$, and $b_0 + b_1 + \dots + b_s = 1$? If the functions ρ_i are known to be continuous on the entire simplex, then it can be shown that they must be linear for any local forecaster.

Suppose we now say that Forecaster A *dominates* Forecaster B *on the outcome i* if the marginal distribution of the ith prediction component for forecaster A given that outcome i occurs is stochastically larger than the corresponding marginal distribution for B. We know from Theorem 7 of Section 3 that, if multivariate forecasters A and B are both well-calibrated and A dominates B on all s outcomes, then A is marginally at least as refined as B with respect to each possible outcome. An open question is whether it is possible to use calibration, locality, linearity as in (5.9), or semi-calibration in connection with some version(s) of dominance to imply that one forecaster is sufficient for (or more refined than) another in our full multivariate sense.

REFERENCES

Blackwell, D. (1951). Comparison of experiments. *Proc. Second Berkeley Symp. Math. Statist. Probability*. Berkeley: University of California Press, 93-102.

Blackwell, D. (1953). Equivalent comparison of experiments. *Ann. Math. Statist.* **24**, 265-272.

Blackwell, D. and Girshick, M.A. (1954). *Theory of Games and Statistical Decisions*. New York: Wiley.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1-3.

Dawid, A.P. (1982). The well-calibrated Bayesian. *J. Amer. Statist. Assoc.* 77, 605-610.

de Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. *Studies in Subjective Probability* (1964). (H.E. Kyburg and H.E. Smokler, eds.), John Wiley, New York.

de Finetti, B. (1962). Does it make sense to speak of 'Good Probability Appraisers'? *The Scientist Speculates -- An Anthology of Partly-Baked Ideas*. (I.J. Good, gen. ed.), Basic Books, New York.

de Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *Brit. J. Math. Statist. Psych.* **18**, 87-123.

DeGroot, M.H. and Eriksson, E.A. (1983). Probability forecasting, stochastic dominance, and the Lorenz curve. Department of Statistics Technical Report No. 291, Carnegie-Mellon University.

DeGroot, M.H. and Fienberg, S.E. (1982). Assessing probability assessors: calibration and refinement. *Statistical Decision Theory and Related Topics III, Vol. 1* (S.S. Gupta and J.O. Berger, eds.), New York: Academic Press, 291-314.

DeGroot, M.H. and Fienberg, S.E (1983). The comparison and evaluation of forecasters. *The Statistician* **32**, 12-22.

Fishburn, P.C. and Vickson, R.G. (1978). Theoretical foundations of stochastic dominance. *Stochastic Dominance* (G.A. Whitmore and M.C. Findley, eds.), Lexington, Mass.: Lexington Books, 39-113.

Fischhoff, B. and MacGregor, D. (1982). Subjective confidence in forecasts. *J. of Forecasting* **1**, 155-172.

Fishburn, P.C. and Vickson, R.G. (1978). Theoretical foundations of stochastic dominance. *Stochastic Dominance* (G.A. Whitmore and M.C. Findley, eds.), Lexington, Mass.: Lexington Books, 39-113.

Good, I.J. (1952). Rational decisions. *J. Roy. Statist. Soc. B* **14**, 107-114.

Habbema, J.D., Hilden, J., and Bjerregaard, B. (1978). The measurement of performance in probabilistic diagnosis. I. The problem, descriptive tools, and measures based on classification matrices. *Methods of Information in Medicine* **17**, 217-226.

Hardy, G.H., Littlewood, J.E., and Polya, G. (1929). Some simple inequalities satisfied by convex functions. *Messenger Math.* **58**, 145-152.

Hardy, G.H., Littlewood, J.E., and Polya, G. (1934). *Inequalities.* Cambridge: Cambridge University Press.

Hilden, J., Habbema, D.F. and Bjerregaard, B. (1978a). The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. *Methods of Information in Medicine* **17**, 226-237.

Hilden, J., Habbema, D.F. and Bjerregaard, B. (1978b). The measurement of performance in probabilistic diagnosis. II. Methods based on continuous functions of the diagnostic probabilities. *Methods of Information in Medicine* **17**, 238-246.

Lindley, D.V. (1980). Scoring rules and the inevitability of probability (with discussion). *Int. Statist. Rev.* **50**, 1-26.

Lindley, D.V., Tversky, A. and Brown, R.V. (1979). On the reconciliation of probability assessments. *J. Roy. Statist. Soc. A* **142**, 146-180.

Marshall, A.W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and its Applications.* New York: Academic Press.

Murphy, A.H. and Winkler, R.L. (1974). Subjective probability forecasts in meteorology: some preliminary results. *Bull. Amer. Meteorol. Soc.* **55**, 1206-1216.

Murphy, A.H. and Winkler, R.L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Appl. Statist.* **26**, 41-47.

Pratt, J.W. (1962). Must subjective probabilities be realized as relative frequencies? Unpublished seminar paper, Harvard University Grad. School of Bus. Administration.

Rothschild, M. and Stiglitz, J.E. (1970). Increasing risk: I. A definition. *J. Econ. Theory* **2**, 225-243.

Rothschild, M. and Stiglitz, J.E. (1973). Some further results on the measurement of inequality. *J. Econ. Theory* **6**, 188-204.

Schervish, M.J. (1983). A general method for comparing probability assessors. Department of Statistics Technical Report No. 275, Carnegie-Mellon University.

Sherman, S. (1951). On a theorem of Hardy, Littlewood, Polya, and Blackwell. *Proc. Nat. Acad. Sci.* **37**, 826-831.

Staël von Holstein, C.A.S. (1970). *Assessment and Evaluation of Subjective Probability Distributions.* Stockholm: Economic Research Institute, Stockholm School of Economics.

Strassen, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist.* 36, 423-439.

Vardeman, S. and Meeden, G. (1983). Calibration, sufficiency and domination considerations for Bayesian probability assessors. Unpublished manuscript, Iowa State University.

Winkler, R.L. (1982). On "good probability appraisers." Unpublished manuscript, Indiana University.

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report No. 306	2. GOVT ACCESSION NO. AD-A135 966	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Comparing Probability Forecasters: Basic Binary Concepts and Multivariate Extensions		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Morris H. DeGroot Stephen E. Fienberg		8. CONTRACT OR GRANT NUMBER(s) N00014-80-0637
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Carnegie-Mellon University Pittsburgh, PA 15213		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Contracts Office Carnegie-Mellon University Pittsburgh, PA 15213		12. REPORT DATE September, 1983 13. NUMBER OF PAGES 22
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)